Hypertexte et manuscrits

Matthieu Bonicel

Matthieu Bonicel

Hypertexte et manuscrits : le défi de l'interopérabilité

BnF. Manuscrits. Français 1586, f. 12 verso

Dans le domaine des manuscrits, les bibliothécaires ont toujours dû jongler avec plusieurs strates de textes, sur et autour du document lui-même. Objet complexe, pouvant à lui seul faire l'objet d'une étude détaillée, le manuscrit est entouré de multiples données textuelles, d'ordre et de valeur diverses : notices, données bibliographiques, catalogues d'exposition, éditions critiques, cartels d'exposition, articles de revues, thèses et mémoires… Cette jungle textuelle donne souvent le vertige au chercheur qui entreprend une étude et qui vient à la rencontre de l'institution de conservation de l'original, non seulement pour accéder au document, mais aussi pour savoir « ce qui a déjà été fait » sur telle ou telle cote.

L'arrivée du numérique n'a rien changé, sinon qu'elle est venue rajouter au-dessus de cet amoncellement une nouvelle couche de texte, celle des fameuses « métadonnées », qui renseignent à la fois les internautes et les machines sur la façon de consulter et d'interagir avec la reproduction virtuelle : fichiers de structure, données encapsulées, notices reproduites dans les formats les plus divers. Et tout cela, rappelons-le, autour d'un ensemble d'images qui, techniquement, ne comportent aucun texte. Car, à la différence fondamentale de l'imprimé, le manuscrit ne s'est pour l'instant affiché sur la toile que sous la forme de lots d'images illisibles par les logiciels de reconnaissance optique de caractères, malgré quelques projets en cours qui nous l'espérons porteront bientôt leurs fruits. L'image restant muette, il est encore plus nécessaire de l'entourer d'un maximum de données textuelles permettant de la rendre clairement identifiable. Et pourtant, les éditions électroniques de textes issus de manuscrits se multiplient, produites la plupart du temps non pas par les bibliothèques conservant les ouvrages originaux, mais par des chercheurs travaillant pour les universités ou d'autres institutions de recherche.

Faciliter l'exploitation des manuscrits numérisés

La multiplication des projets de numérisation au niveau international a fait apparaître un nouvel enjeu : celui de l'interopérabilité. Beaucoup d'entreprises

désormais célèbres comme l'*International Dunhuang Project* [1] ou *Europeana Regia* [2] se sont revendiquées comme des reconstitutions virtuelles de collections physiquement dispersées, afin de faciliter les travaux de comparaison, qui ont toujours été la base des études sur les manuscrits. Cependant, les moyens techniques mis en œuvre n'ont à ce jour pas permis de répondre à une demande essentielle des chercheurs : la possibilité de comparer automatiquement en un lieu unique plusieurs manuscrits numérisés mis en ligne par des institutions de conservation différentes. Certains portails, comme la remarquable *Roman de la Rose Digital Library* [3], permettent bien de visualiser dans une même interface des manuscrits issus de bibliothèques différentes, mais leur réalisation a nécessité la copie manuelle des images en question sur un serveur unique, sans qu'il soit possible de récupérer les images à la demande sur la bibliothèque numérique d'origine (*Gallica* par exemple en ce qui concerne la France). L'internaute qui, au détour de ses recherches sur la toile, découvre deux manuscrits qu'il souhaite comparer sur deux sites d'institutions différentes n'a pas la possibilité de les afficher dans un outil unique et doit se débrouiller en circulant d'un onglet à l'autre de son navigateur. De même, si à l'issue de son travail à partir des images numériques il met en ligne une édition électronique normalisée du texte conservé dans ces deux manuscrits, il n'a pas la possibilité de rendre celui-ci accessible à partir des bibliothèques numériques qui conservent les images des documents originaux.

Bien sûr, des liens hypertextes sont toujours possibles, et c'est d'ailleurs une pratique de plus en plus répandue dans les bibliothèques, qui enrichissent souvent leur catalogue de références à des travaux accessibles en ligne sur le site de leurs partenaires universitaires. Depuis deux ans, cependant, un groupe d'experts internationaux, réunis à l'université Stanford sous le patronage de la Fondation Andrew W. Mellon, qui a déjà financé de nombreux projets de ce genre, s'est penché sur la question de l'interopérabilité dans le domaine des bibliothèques numériques de manuscrits, en se demandant si l'on ne pourrait pas franchir le cap décisif d'une interaction automatisée entre les différents outils et entrepôts disponibles. Cette équipe est constituée d'une part de représentants d'institutions mettant en ligne des manuscrits numérisés comme la BNF, la British Library à Londres, la Bodleian Library à Oxford, le portail suisse *E-codices* [4], l'université Stanford avec *Parker on the Web* [5] ou la Johns Hopkins University de Baltimore avec son portail sur le *Roman de la rose* et d'autres part de développeurs ayant mis en place des outils d'exploitation des images numériques de

1 Numérisation et mise en ligne des manuscrits et objets d'arts découverts au début du xxe siècle à Dunhuang en Chine et sur la route de la Soie et aujourd'hui dispersés à travers le Monde dans différentes institutions : <http://idp.bnf.fr> (consulté le 29 août 2012).

2 Numérisation afin d'alimenter le portail Europeana d'un millier manuscrits conservés dans six grandes bibliothèques européennes et provenant de trois grandes collections royales : les manuscrits carolingiens, la Librairie de Charles V et la bibliothèque des rois aragonais de Naples : <http://www.europeanaregia.eu> (consulté le 29 août 2012).

3 Accès à un grand nombre de manuscrits du Roman de la rose, dont les 140 manuscrits conservés en France, sur un site développé par la Johns Hopkins University de Baltimore aux États-Unis : <http://romandelarose.org> (consulté le 29 août 2012).

4 Projet de numérisation et de mise en ligne des manuscrits conservés dans diverses institutions suisses piloté par l'université de Fribourg : <http://www.e-codices.unifr.ch> (consulté le 29 août 2012).

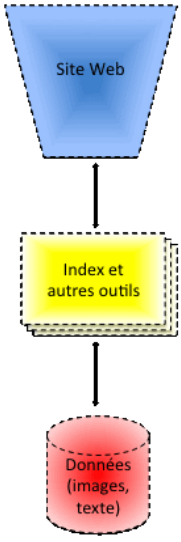
5 Bibliothèque virtuelle des manuscrits du Corpus Christi College de Cambridge : <http://parkerweb.stanford.edu> (consulté le 29 août 2012).

manuscrits, comme le projet T-PEN [6] de la Saint Louis University ou Digital Mappaemundi [7] de Drew University.

Dès les premières réunions, les représentants des institutions de conservation ont souligné un besoin urgent de trouver une solution permettant une forte interopérabilité avec les équipes porteuses de projets, afin d'éviter de multiplier les copies récurrentes de documents numériques, ce qui avait pour effet d'alourdir considérablement les procédures et de multiplier les points de stockage et de sauvegarde, sans véritable possibilité de mise à jour automatique ou de contrôle de la façon dont les données étaient réutilisées. De leur côté, les développeurs chargés de projets disposaient souvent de ressources limitées et d'un calendrier serré, et le temps passé à manipuler d'importants lots de données les détournait de leur tâche principale d'exploitation de ces données. Il est donc rapidement apparu nécessaire de construire un modèle de données permettant de décrire l'existant afin de le rendre pleinement accessible aux différents utilisateurs, et donc d'introduire une couche supérieure de texte au-dessus des textes existants.

Vers un nouveau modèle de données : du « silo de contenus » aux « canevas partagés »

Pour comprendre la façon dont le groupe d'experts a élaboré ce modèle de données, il est nécessaire d'appréhender la façon dont la plupart des bibliothèques numériques sont aujourd'hui conçues et les enjeux techniques que présente la mise en interopérabilité des systèmes existants. À l'heure actuelle, la plupart des sites Internet mettant en ligne des fac-similés numériques de manuscrits fonctionnent selon la logique du « silo de contenus » : un entrepôt de données (images et métadonnées) sur lequel on vient placer des index et d'autres outils (panier de données, annotation, etc.), le tout coiffé d'un site Web, vitrine publique du silo comprenant un visualiseur d'images, plus ou moins élaboré.



Les fonctionnalités offertes peuvent être assez évoluées, mais se concentrent toujours sur les documents stockés localement, à savoir dans l'entrepôt sur lequel les outils sont assis. Il n'existe pas encore d'entreprise de grande ampleur permettant d'afficher dans le contexte graphique et applicatif d'une bibliothèque numérique des documents stockés dans l'entrepôt d'une autre bibliothèque numérique. Ainsi, il n'est pas possible d'afficher dans *Gallica* les images d'un document numérique mise en ligne par la British Library, et *vice versa*. La comparaison sur un même écran de deux documents numériques conservés dans deux espaces de stockage différents est donc impossible non pas parce qu'il n'existe pas encore de visualiseur compatible mais parce que la logique de stockage des entrepôts

6 Outil d'assistance à la transcription de manuscrits médiévaux permettant un découpage automatique des images par repérage des lignes d'écriture : <http://www.t-pen.org> (consulté le 29 août 2012).

7 Outil d'annotation collaborative d'image et de textes : <http://ada.drew.edu/dmproject> (consulté le 29 août 2012).

numériques ne permet pas encore le partage d'images à la volée, au gré des internautes. Il est vrai en revanche que la mise en place de plus en plus systématique d'entrepôts d'exposition des métadonnées selon le protocole OAI-PMH [8] permet d'afficher les métadonnées concernant les documents originaux ou leur reproduction numérique dans un contexte différent. C'est grâce à ce protocole que *Gallica*, au niveau national, et *Europeana* ou *The European Library* par exemple au niveau européen, peuvent donner accès aux notices d'un très grand nombre de documents numériques mis en ligne par des institutions différentes. Cependant, ce mode d'interopérabilité se limite aux métadonnées descriptives et à une vignette illustrative, et l'utilisateur est renvoyé au site local pour consulter le document numérique dans son intégralité. Il n'a donc pas le choix du mode de visualisation et des outils qui sont associés au document numérique.

Le principe de l'interopérabilité vise plutôt à déconnecter le document numérique et ses métadonnées primaires (techniques et bibliographiques) des outils et modes de visualisation qu'on peut lui appliquer. Dans ce schéma, les bibliothèques numériques, les chercheurs et les logiciels destinés à faciliter la recherche constituent trois entités indépendantes les unes des autres mais qui peuvent communiquer comme bon leur semble. En d'autres termes, un chercheur doit pouvoir être libre de choisir les documents auxquels il souhaite accéder de même que les outils avec lesquels il souhaite les exploiter, sans être tributaire ni des uns ni des autres. De même, les données qu'il produit sur les documents numériques à l'aide des outils qui lui sont proposés doivent rester accessibles, dans un format normalisé, et ne pas être cantonnées à l'affichage dans le site de l'outil avec lequel elles ont été produites. L'interopérabilité, c'est donc aussi assurer la pérennité des nouvelles données produites, indépendamment des outils qui ont permis leur production.

Les manuscrits médiévaux étant des objets complexes, dans leur forme et leur structure primaire mais aussi dans la forme finale que l'historique de leur conservation leur a donnée, il est nécessaire pour décrire leur version numérique de remonter à un certain niveau d'abstraction. En effet, le document numérique issu de la dématérialisation de l'original est quant à lui un ensemble simple composé d'une séquence d'images. Le contenu de ces images en revanche, la façon de les afficher et de décrire ce qu'elles représentent est beaucoup plus complexe. L'image 1 par exemple ne correspond quasiment jamais au folio ou à la page 1 du manuscrit, car il s'agit souvent de l'image du plat supérieur de la reliure. Le folio 1 se situe souvent aux alentours de l'image 10, après la reliure, les gardes et les éventuels feuillets liminaires non numérotés. De même, il existe parfois des documents numériques parallèles ne comportant que certaines vues du manuscrit, celles des enluminures ou des colophons photographiés sous lumière ultraviolette. En l'absence du document physique, il convient donc d'indiquer au lecteur où dans l'organisation du document original se situent les vues qu'il est en train de consulter. Évidemment, nous n'évoquons ici que des cas connus depuis longtemps et que chaque institution a pris en compte dans la construction de sa bibliothèque numérique. À la BNF par exemple, un fichier XML [9] de

8 Open Archives Initiative – Protocol for Metadata Harvesting. On peut se référer à la page décrivant sommairement le protocole sur le site de la BNF :
<http://www.bnf.fr/fr/professionnels/protocoles_echange_donnees/a.proto_oai.html> (consulté le 29 août 2012).

9 eXtensible Markup Language ou « langage de balisage extensible ». Il s'agit d'un langage de programmation utilisant le principe des balises ouvrantes et fermantes (comme HTML ou SGML) mais permettant de définir soi-même, de manière plus ou moins normalisée, le nom des balises elles-mêmes. Il permet donc une grande flexibilité d'utilisation, dans les domaines d'applications les plus divers. C'est un des moyens les plus répandus aujourd'hui de structurer des données de manière hiérarchisées, les éléments du fichier de données étant emboîtables comme des

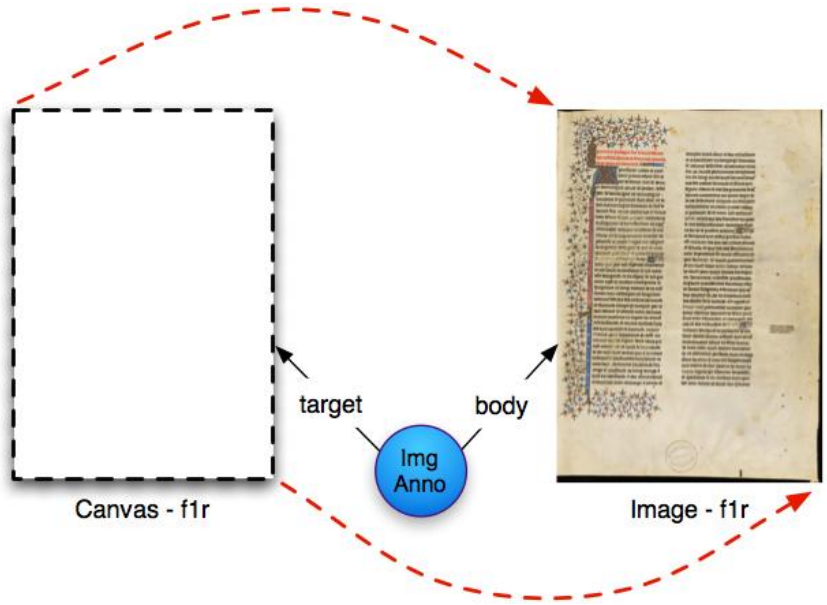
métadonnées, nommé RefNum, donne la concordance entre l'ordre des images et celui des feuillets du manuscrit. Mais la tâche se complique lorsque l'on doit mettre en relation les images d'un document numérique et les autres instances digitales produites par des institutions diverses : images, éditions électroniques, bases de données, bibliographie, sites Internet dédiés à une œuvre… La nécessité d'un référentiel commun, dont les modalités d'expression sont connues de tous au niveau international, permet d'être sûr de faire référence à la bonne page du manuscrit original ou à la bonne ligne de l'édition électronique.

C'est l'image du canevas qui a été retenue, et plus précisément celle d'un empilement de différentes toiles vierges au sein duquel les différents contenus numériques viennent s'entremêler. De là est né le nom du modèle de données : SharedCanvas, les « canevas partagés ».

Principes du nouveau modèle de données

La création d'une couche d'abstraction

Le fonctionnement du modèle repose sur un principe simple. La création d'une entité abstraite représentant l'unité physique de base (la page, le feuillet, le fragment) d'un côté et la ressource se rapportant à cette entité abstraite de l'autre. Le lien entre ces deux éléments se nomme une annotation, la première d'entre elles étant souvent l'image numérique représentant le document physique. L'annotation pointe d'un côté vers le canevas, entité abstraite, en tant que cible (*target*) afin de se référencer dans cet élément centralisateur et de l'autre vers l'image numérique elle-même en tant que corps (*body*) de l'annotation qu'elle représente.



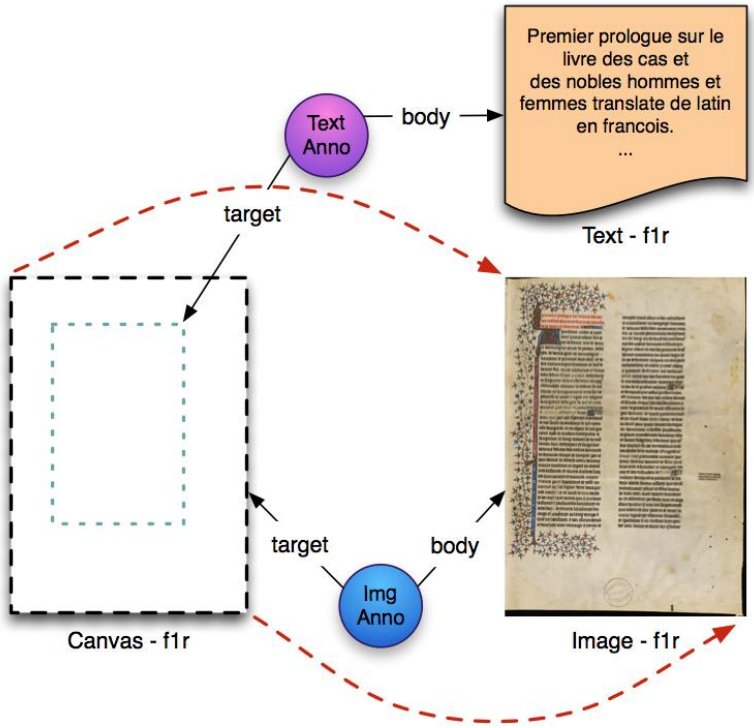
L'unité de base : le canevas

poupées gigognes.

Le modèle SharedCanvas institue comme élément de base un canevas, unité d'abstraction se rapportant à l'unité de base de l'objet physique décrit. Dans le cadre d'un manuscrit en volume, le canevas correspondra dans la très grande majorité des cas à la face d'un feuillet, recto ou verso. Mais un canevas pourrait également identifier la face d'un rouleau, ou un fragment de papyrus par exemple. Destiné à être le point de référence entre le document numérique et les différentes annotations ou informations s'y rapportant, le canevas ne pointe pas vers l'objet physique lui-même mais vers sa représentation numérique. De forme rectangulaire, il s'identifie donc par une largeur et une hauteur calculées entre le coin supérieur gauche et le coin inférieur droit de l'image numérique auquel il se rapporte, comme le symbolisent les flèches rouges de l'illustration 2. Il comprend également un élément, *label*, permettant de l'identifier de manière claire (en général il s'agira du numéro de feuillet) et, enfin, une liste d'annotations qui lui sont associées.

Les annotations textuelles

Une fois la taille de base du canevas définie, toute annotation venant s'appuyer sur une partie du feuillet du manuscrit (transcription d'un passage par exemple) va donc pointer d'un côté vers une zone du canevas et de l'autre vers le texte de la transcription proposée. Ce qu'il faut bien comprendre, c'est qu'il n'y a alors plus de relation directe entre l'annotation textuelle et l'image du document numérique. Cette relation est déportée dans le canevas, qui sert d'abstraction commune afin de ne pas rendre tel ou tel élément dépendant d'un autre. Si pour une raison ou pour une autre l'image numérique est supprimée, les annotations et leur localisation subsisteront. Il faudra toutefois que l'image venant remplacer celle qui a disparue sache gérer les correspondances entre l'ancienne et la nouvelle représentation de l'objet physique.



La gestion de multiples séquences

Là où la plupart des fichiers de structure pour bibliothèque numérique se basent sur une simple séquence continue d'images (du premier au dernier feuillet, avec ou sans la reliure), le modèle de données mis en place permet de prendre en compte un grand nombre de cas de figures, voire la mise en place de séquences parallèles. Il doit par exemple être possible d'exprimer clairement le fait que les feuillets d'un manuscrit comportent plusieurs foliotations concurrentes, apposées à différentes périodes, et ne respectant pas forcément toutes le même ordre, certaines bibliothèques ayant à certaines époques pris le parti de renuméroter les feuillets de leur ouvrage pour restituer un ordre supposé correct et perturbé par une erreur de reliure. Il est également possible d'exprimer le fait qu'un même texte, composé d'une séquence cohérente de feuillets, se retrouve, du fait de son volume très important, relié en trois volumes distincts. *A contrario*, un même objet codicologique peut comporter plusieurs séquences successives d'œuvres différentes. Enfin, pour ne donner qu'un autre exemple de l'infini des combinaisons possibles, différents feuillets d'un manuscrit peuvent se révéler être des fragments tirés d'un autre volume et insérés autre part pour des raisons historiques diverses. Le fichier XML de séquence permet de répondre à toutes ces exigences et d'établir le lien entre les canevas de chaque unité de base et les images numériques correspondantes. Il peut également être utilisé pour fournir les informations nécessaires à un visualiseur pour présenter les images au bon format, dans le bon sens (de droite à gauche ou de gauche à droite dans le cas d'un livre oriental à présenter en deux pages par image par exemple) et avec les informations utiles à l'affichage (le numéro des feuillets et non celui de l'image dans la séquence par exemple).

Le manifeste, expression des différentes instances

L'implantation du modèle SharedCanvas suppose la mise en place, sur un serveur accessible par les partenaires potentiels, que ce soit en libre accès total ou par le biais d'une accréditation, de fichiers XML RDF[10] présentant différents éléments. Le manifeste, fichier de base, donne une identification sommaire de chaque document, l'institution responsable de sa mise en ligne et la liste des autres fichiers disponibles correspondant au document. On y trouve notamment la référence de la séquence et de la liste des annotations, qui correspondent chacun à un autre fichier RDF, voire plusieurs, s'il existe par exemple plusieurs séquences possibles pour un même document.

Lorsqu'une autre institution désirant utiliser SharedCanvas réalise un travail sur un manuscrit numérique, elle crée sur son propre serveur une nouvelle série de fichiers, à commencer par un manifeste qui fait référence au manifeste de l'institution hébergeant le fac-similé numérique. Par un jeu de moissonnage réciproque, chaque institution peut alors inclure dans son manifeste la référence au manifeste des autres partenaires et donner ainsi une liste aussi complète que possible des ressources disponibles sur un même document physique. Par exemple, il est possible d'associer les images d'un manuscrit numérisé par une bibliothèque au texte du manuscrit édité par une université. Chacune des deux institutions met en place un manifeste qui signale automatiquement et de manière dynamique ses propres ressources et celles de ses partenaires.

10 eXtensible Markup Language (cf. note n° 9) Ressource Description Framework : développé par le W3C, RDF est un modèle de données destiné à décrire de faon formelle les ressources du Web sémantique et les métadonnées s'y rapportant. <http://www.w3.org/RDF/> (consulté le 29 août 2012)

Au cours de ses deux premières années de travaux, l'équipe du projet SharedCanvas a constitué un site Internet [11] sur lequel on peut trouver à la fois la documentation du modèle de données et des premiers tests d'implémentation. La BNF, de son côté, a participé dans le cadre d'un projet de numérisation des manuscrits de Guillaume de Machaut, à la mise en place, sur un serveur de test à l'université Stanford, de documents numériques disponibles pour une exploitation à la volée dans le logiciel T-PEN. Les utilisateurs test peuvent, à partir de l'interface de T-PEN, se faire envoyer les images directement depuis Stanford, procéder au découpage automatique suivant les lignes d'écriture et commencer leur transcription en quelques secondes. L'étape suivante sera la mise en place effective des instances SharedCanvas dans les différentes institutions partenaires. Un pack de développement est actuellement en cours de finalisation pour permettre à chaque établissement qui le souhaite de mettre rapidement en place une chaîne applicative permettant la production de fichiers SharedCanvas et ainsi rendre son contenu interopérable. Le travail n'est pour autant pas terminé car il reste encore beaucoup de questions techniques à résoudre, ainsi que des questions d'ordre juridique, concernant par exemple la politique de propriété intellectuelle à appliquer lorsque des contenus pourront librement, ou sur authentification, circuler automatiquement sur le Web et être manipulés par des outils tiers. Cette refléxion devrait s'enrichir dans les années à venir avec le développement du *cloud computing* où de plus en plus de données seront disponibles dans le nuage à quiconque saura les appréhender et disposera d'un droit d'accès, indépendamment de leur stockage physique par telle ou telle institution.

11 <http://www.shared-canvas.org> (consulté le 29 août 2012).